Machine Learning Algorithms in Healthcare: Disease Prediction

Arpit Dua
Software Engineering
Delhi Technological University
Delhi,India

arpitdua_se20b8_17@dtu.ac.in

Abstract— **This paper presents a novel approach for disease prediction using machine learning algorithms. The proposed method leverages machine learning techniques to analyze medical records and to predict likelihood of diseases in future. We demonstrate the effectiveness of our approach using the " Cardiovascular Diseases Risk Prediction Dataset " From the Kaggle website. Experimental results show that the certain algorithms offer advantages over the other in specific diseases**.

Keywords— **Machine learning in Healthcare, Disease prediction, Machine Learning, Machine learning algorithms**

I. Introduction

The incorporation of Machine Learning (ML) into clinical practice signifies a significant shift with promising prospects for improving healthcare delivery. Increasingly, private healthcare organizations are acknowledging the potential of ML in decision-making. Customized ML algorithms designed for specific functions within the medical sector are gaining prominence. Researchers are exploring the use of ML tools for big data analysis across various medical disciplines, indicating a transformative trajectory for the industry [9] [13]. ML algorithms are now being created with distinct roles in diverse medical domains, reflecting a shift towards algorithmic autonomy.

Notably, in fields such as radiology and anatomical pathology, these algorithms are outperforming traditional diagnostic methods. The healthcare sector is witnessing a surge in the adoption of ML tools for scrutinizing medical images, thereby augmenting the intelligence of healthcare systems. It's crucial to note that the integration of ML aims to enhance the intelligence of the entire healthcare ecosystem rather than replacing physicians [22].

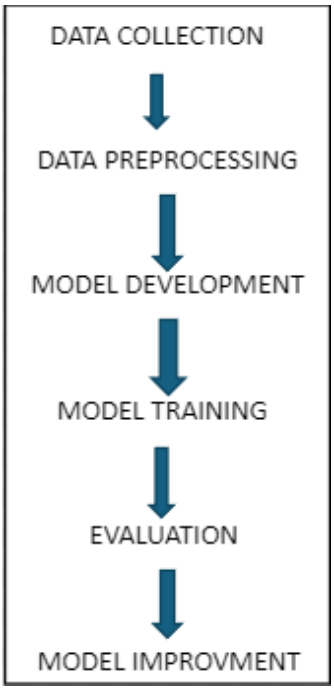


Figure 1: Machine Learning Process which depicts data collection from EHR for Model Training

II. Problems

The rising cost of healthcare presents a need for Machine learning intervention for a better healthcare delivery for all patients across the world. Machine learning is not there to replace doctors but instead help better and faster diagnosis of patients. However, Machine learning algorithms are still a long way from being feasible for actual use. Improvement in accuracy as well as established minimum confidence level should be present for practical use.

III. Methodology

This paper presents the accuracy of machine learning algorithms in Healthcare, which is achieved by the following tools.

A. *Weka tool*

The research made use of WEKA tool as it helps in performance evaluation and performing comparison of various machine learning algorithms on dataset without any coding required. The WEKA tool is briefly described below:
WEKA [Waikata Enviroment for Knowledge Analysis]
• It is a very popular machine learning and data mining toolkit for conducting data driven research.
• Developed in University of Waikato, New Zealand
• All algorithms are written in Java
• The version of WEKA used for experimentation in this paper is WEKA Version 3.8.6

B. *CARDIOVASCULAR DISEASE PREDICTION DATASET*

Dataset used in our research is obtained from Kaggle. The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Dataset is already preprocessed. Originally it had 304 variables which are reduced to 19 variables which could directly affect Cardiovascular health. We experiment with four machine learning algorithms, including Logistic Regression, Random Forest, Naive Bayes, and k-Nearest Neighbors (KNN). Finally, these trained models are deployed to predict whether a patient is likely to have heart disease based on his lifestyle choices and age.

EXPERIMENTAL SETUP

The experiments are conducted using a dataset obtained from real-world smart home environments. The dataset comprises various sensor readings, including temperature, humidity, motion, & energy consumption. Each algorithm is implemented and evaluated using Python programming language with appropriate libraries such as scikit-learn.

A. CONFUSION MATRIX

CM (Confusion Matrix) is used to check for accuracy,sensitivity,specificity,etc of a model made using Machine learning. This Matrix is used in almost all evaluation-based papers and hence is necessary to understand before we conduct the Literature review.

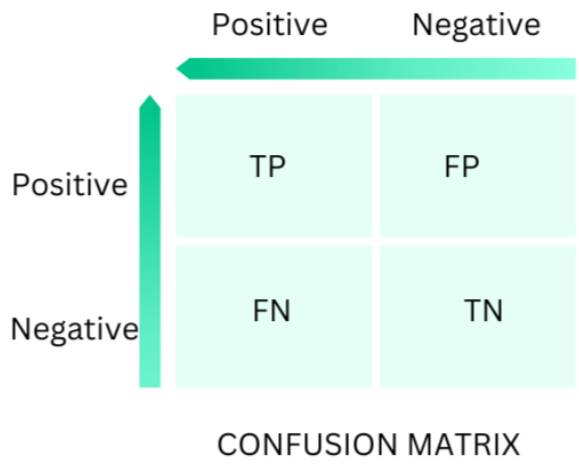


Figure 2: CM where the rows are predicted values and columns are actual values

There are only 4 types of outcomes possible based on actual vs Predicted values that are: -

1) True Positives (TP)

Actual Positive and Predicted Positive

2) True Negatives (TN)

Actual Negative and Predicted Negative

3) False Positives (FP)

Actual Negative and Predicted Positive

4) False Negatives (FN)

Actual Positive and Predicted Negative

Above parameters can be used to calculate various Evaluation metrics as follows:

- Accuracy

Accuracy can be calculated using true values upon total sample population.it signifies the correct predictions of a model

Accuracy = (TP+TN)/(TP+TN+FP+FN)

- Precision

Precision is calculated as true positives in a model upon total positives. Signifies how well a model can Predict TP of a class.

Precision =TP/(TP+FP)

- Sensitivity or Recall

Sensitivity or Recall is calculated by dividing True positives by total Actual positives (FN is actually positive). Measures how well a model can predict Positive instances

Recall = TP/(TP+FN)

B. ALGORITHMS IMPLEMENTED

We implement several machine learning algorithms for disease prediction. Understanding how each of these algorithm's work is necessary to get their current contribution and future improvements in the field of healthcare. The algorithms used in this research are explained in brief below:

- Naive Bayes- Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem which assumes independence between features.

Naive Bayes can perform well in many complex real-world scenarios. It is very useful when dealing with high-dimensional datasets and is often used in text classification (spam detection, sentiment analysis), recommendation systems, medical prognosis and medical diagnosis.

- Random Forest- Random Forest is an ensemble learning method. It constructs many decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Random Forest is known for its robustness, scalability, and ability to handle big data with complex interactions between features. It is useful in classification, regression, and feature selection.

- Logistic Regression- Logistic regression is a linear model for binary classification that uses the logistic function (sigmoid function) to model the probability that a given input belongs to a particular class. Logistic regression is simple, and efficient, making it particularly useful for binary classification tasks. It's widely used in various fields such as healthcare, marketing, and finance.

- KNN- KNN is a non-parametric classification algorithm that classifies new data points based on the majority class among their k nearest neighbors in the feature space. KNN is simple to understand and implement, but it can be computationally expensive, especially with large datasets. It's commonly used in recommendation systems, pattern recognition, and anomaly detection.

RESULTS

The performance metrics of the Disease prediction applied to the Cardiovascular health dataset are summarized in the table below:

TABLE I
PERFORMANCE METRICS OF DISEASE PREDICTION ALGORITHMS

Algorithm	Accuracy (%)	Precision	Recall	F-measure
Naive Bayes	88.36	0.311	0.338	0.324
KNN	86.85	0.206	0.207	0.206
Logistic Regression	91.72	0.489	0.057	0.103
Random Fores	91.55	0.405	0.053	0.094

A. CONFUSION MATRICES

Fig. 3. K-Nearest Neighbour

```
=== Confusion Matrix ===

     a     b   <-- classified as
 52590  4078 |     a = No
  4048  1055 |     b = Yes
```

Fig. 4. Logistic regression

```
=== Confusion Matrix ===

     a     b   <-- classified as
 56362   306 |     a = No
  4810   293 |     b = Yes
```

Fig. 5. Random Forest

```
=== Confusion Matrix ===

     a     b   <-- classified as
 56270   398 |     a = No
  4832   271 |     b = Yes
```

Fig. 6. Naive Bayes

```
=== Confusion Matrix ===

     a     b   <-- classified as
 52856  3812 |     a = No
  3380  1723 |     b = Yes
```

Figure 3, Figure 4, Figure 5 and Figure 6 represent confusion matrices we obtained using the WEKA tool. The results of these matrices were used to calculate the measures in Table 1.

B. PLOT MATRIX WEKA

We can also visualize the data using Weka tools. The dataset attributes are marked on the x-axis and y-axis while the instances are plotted. Plot Matrices can help identify patterns and predict without models.
In our research, we have used heart disease attribute as the color while plotting different factors on x-axis and y-axis.

Fig. 7. Plot Matrix (X:Age; Y:FriedConsumption)

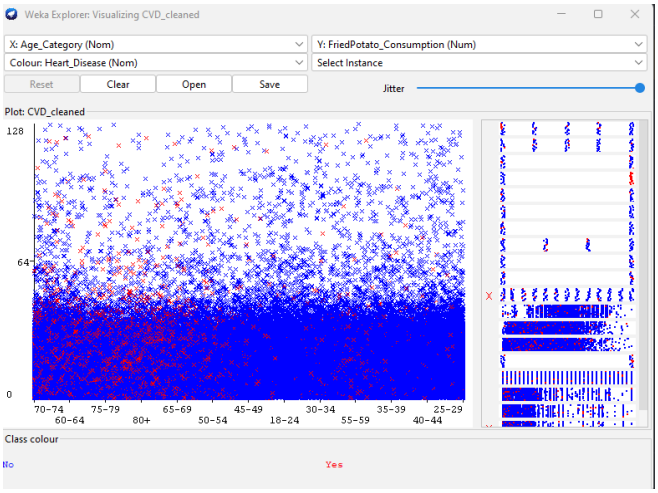
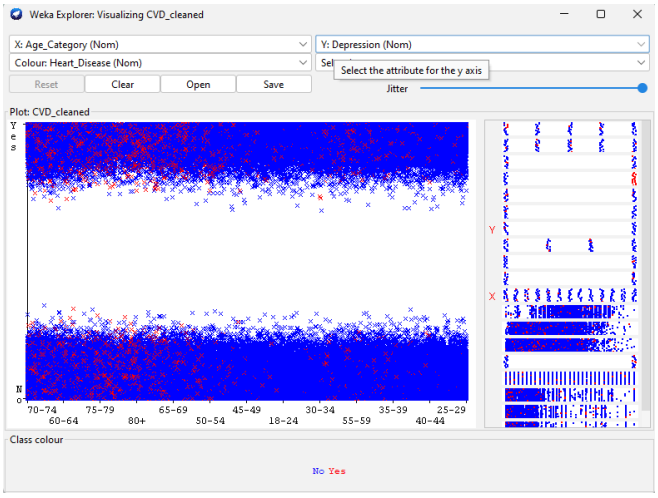


Fig. 8. Plot Matrix (X:Age; Y:Depression)



In figure 7 and figure 8, the red dot represents the patients that were diagnosed with cardiovascular diseases and blue dots that were not. Their lifestyle choices were plotted along with their age. As per Figure 7, People consuming fried food after the age of 60 were more likely to have heart disease as compared to people below 60. As per Figure 8, People who had depression are more likely to have heart disease compared to the others of same age group.

REFERENCES

(1) Callahan, Alison, and Nigam H. Shah. "Machine learning in healthcare." *Key advances in clinical informatics*. Academic Press, 2017. 279-291.

(2) Bhardwaj, Rohan, Ankita R. Nambiar, and Debojyoti Dutta. "A study of machine learning in healthcare." *2017 IEEE 41st annual computer software and applications conference (COMPSAC)*. Vol. 2. IEEE, 2017.

(3) Qayyum, Adnan, et al. "Secure and robust machine learning for healthcare: A survey." *IEEE Reviews in Biomedical Engineering* 14 (2020): 156-180.

(4) Ferdous, Munira, Jui Debnath, and Narayan Ranjan Chakraborty. "Machine learning algorithms in healthcare: A literature survey." *2020 11th International conference on computing, communication and networking technologies (ICCCNT)*. IEEE, 2020.

(5) Waring, Jonathan, Charlotta Lindvall, and Renato Umeton. "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare." *Artificial intelligence in medicine* 104 (2020): 101822.

(6) Dalal, Kushal Rashmikant. "Analysing the implementation of machine learning in healthcare." *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 2020.

(7) Maes, Frederik, et al. "The role of medical image computing and machine learning in healthcare." *Artificial intelligence in medical imaging: opportunities, applications and risks* (2019): 9-23.

(8) Rudin, Cynthia, and Berk Ustun. "Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice." *Interfaces* 48.5 (2018): 449-466.

(9) Seneviratne, Martin G., Nigam H. Shah, and Larry Chu. "Bridging the implementation gap of machine learning in healthcare." *Bmj Innovations* 6.2 (2020).

(10) Jain, Vishal, and Jyotir Moy Chatterjee. "Machine learning with health care perspective." *Cham: Springer* (2020): 1-415.

(11) Esteva, Andre, et al. "A guide to deep learning in healthcare." *Nature medicine* 25.1 (2019): 24-29.

(12) Pianykh, Oleg S., et al. "Improving healthcare operations management with machine learning." *Nature Machine Intelligence* 2.5 (2020): 266-273.

(13) Nithya, B., and V. Ilango. "Predictive analytics in health care using machine learning tools and techniques." *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2017.

(14) Gupta, Aakansha, and Rahul Katarya. "Social media based surveillance systems for healthcare using machine learning: a systematic review." *Journal of biomedical informatics* 108 (2020): 103500.

(15) Araújo, Flávio HD, André M. Santana, and Pedro de A. Santos Neto. "Using machine learning to support healthcare professionals in making preauthorisation decisions." *International journal of medical informatics* 94 (2016): 1-7.

(16) Ayanouz, Soufyane, Boudhir Anouar Abdelhakim, and Mohammed Benhmed. "A smart chatbot architecture based NLP and machine learning for health care assistance." *Proceedings of the 3rd*

international conference on networking, information systems & security. 2020.
(17) Farahani, Bahar, Mojtaba Barzegari, and Fereidoon Shams Aliee. "Towards collaborative machine learning driven healthcare internet of things." *Proceedings of the International Conference on Omni-Layer Intelligent Systems*. 2019.
(18) Chen, Po-Hsuan Cameron, Yun Liu, and Lily Peng. "How to develop machine learning models for healthcare." *Nature materials* 18.5 (2019): 410-414.
(19) Chen, Min, et al. "Disease prediction by machine learning over big data from healthcare communities." *Ieee Access* 5 (2017): 8869-8879.
(20) Kempa-Liehr, Andreas W., et al. "Healthcare pathway discovery and probabilistic machine learning." *International journal of medical informatics* 137 (2020): 104087.
(21) Singh, Rameshwer, and Rajeshwar Singh. "Applications of sentiment analysis and machine learning techniques in disease outbreak prediction–A review." *Materials Today: Proceedings* 81 (2023): 1006-1011.
(22) Bhosale, Yogesh H., and K. Sridhar Patnaik. "Application of deep learning techniques in diagnosis of covid-19 (coronavirus): a systematic review." *Neural processing letters* 55.3 (2023): 3551-3603.
(23) Wojtusiak, Janusz. "Semantic data types in machine learning from healthcare data." *2012 11th International Conference on Machine Learning and Applications*. Vol. 1. IEEE, 2012.
(24) Emanet, Nahit, et al. "A comparative analysis of machine learning methods for classification type decision problems in healthcare." *Decision Analytics* 1 (2014): 1-20.
(25) Mir, Ayman, and Sudhir N. Dhage. "Diabetes disease prediction using machine learning on big data of healthcare." *2018 fourth international conference on computing communication control and automation (ICCUBEA)*. IEEE, 2018.
(26) Center for Disease Control, 2021 BRFSS Survey Data and Documentation, https://www.cdc.gov/brfss/annual data/annual 2021.html, 2022.
(27) Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016

PYTHON IMPLEMENTATION:

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.metrics import
accuracy_score,confusion_matrix,recall_score,f1_score
,precision_score

from sklearn.naive_bayes import GaussianNB
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier

import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler

df=pd.read_csv('CVD_cleaned.csv')

df_encoded = df.copy()

#yes and no to 1 and 0;
for col in df_encoded.columns:
    if df_encoded[col].dtype == 'object':
        le = LabelEncoder()
        df_encoded[col] =
le.fit_transform(df_encoded[col])
```

```
#all columns except one
X=df_encoded.drop('Heart_Disease', axis=1)
y=df_encoded['Heart_Disease']

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

X_train, X_test, y_train, y_test =
train_test_split(X_scaled, y, test_size=0.3)

models={"Naive Bayes": GaussianNB(),
        "RF":RandomForestClassifier(n_estimators=100,
class_weight='balanced'),
        "Logistic
reg":LogisticRegression(max_iter=1000,
class_weight='balanced'),
        "KNN":KNeighborsClassifier(n_neighbors=3)}

c_m={}
results=[]

for name,model in models.items():
    model.fit(X_train,y_train)
    prediction=model.predict(X_test)

    accuracy=accuracy_score(y_test,prediction)
```

```
precision=precision_score(y_test,prediction,zero_division=0)

recall=recall_score(y_test,prediction,zero_division=0)
    f1=f1_score(y_test,prediction,zero_division=0)
    cm=confusion_matrix(y_test,prediction)
    c_m[name]=cm

    results.append({
        "Algorithm": name,
        "Accuracy": round(accuracy * 100, 2),
        "Precision": round(precision, 3),
        "Recall": round(recall, 3),
        "F1-Score": round(f1, 3)
    })

results_df=pd.DataFrame(results)
results_df = results_df.sort_values(by="Accuracy", ascending=False)

print("\n Performance Metrics: ")
print(results_df)
```